Dataset : Train_details_22122017.csv.zip

Dataset:

Train schedule across the country including departing station, departure time, next arriving station, train name, and sequence of stations.

Problem Statement:

The station master at every train station is tasked with continuously assessing whether the number of platforms available in that station at any point in time during the day is sufficient for the trains expected to stop at that station. Each station master is looking for help to solve this challenge by having at his/her fingertip the total number of trains at that station in any 20 minute window. It is assumed that a typical train needs about 20 minutes at a station at the maximum for loading & unloading before moving on or being moved to the shed.

Your task is to compute the 20-minute rolling count of trains per station so that every 5 minutes, every station master that has any train stopped at their station gets this count information for their purposes. You are also required to determine per station the maximum of these counts across the entire time period of the data set.

Suggestions:

1. First, convert the batch data into a data stream by loading into Kafka using Spark batch.

2. When doing so, pay attention to the quality of the data (e.g. you will find at times arrival time for the origination station marked as 0:00:00 while at other times it might have the same date as the departure time).

3. Second, write your spark streaming code pointing to that Kafka topic for emitting answers every 5 minutes. While testing your code, you may want to use a smaller time window for streaming (e.g. 10 seconds) so that you are not waiting for the next stream window of 5 minutes.

Notes:

- Demonstrate live outputs
- Do not assume the data file is sorted in ascending order by date
- Do not assume the data has no gaps in dates for each city & region combination.

- Assume that there will be bad rows
- Note that a train that arrives at a station at, say, 9:55 and departs at 10:05, straddles two 20-minute rolling windows – i.e. one corresponding to 9:40 to 10:00 and another from 10:00 to 10:20.

Please Note:

Don't discuss anything with others and whatever you know please do it within the allotted time and submit the code + screenshots of the process and screen recording of the outputs .

Submission should be uploaded to folder Final_Assignment_<dateofsubmission>_<FirstName>